



SMW and Linked Open Data

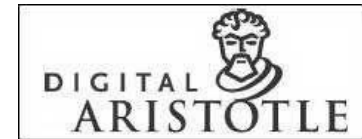
Mark Greaves
Vulcan Inc.

Introduction: Vulcan and SMW

■ What/Who is Vulcan?

■ Vulcan's Interest in SMW

- Not primarily commercial or for internal use
- The Digital Aristotle Vision
 - Hold a vast amount of scientific knowledge
 - Answer questions based on the knowledge
 - Dramatically accelerate scientific progress
- What the Digital Aristotle requires:
 1. Technology to enable a global, widely-authored, very large knowledge base about human affairs and science
 2. Technology that answers questions and proactively supplies information
 3. Technology that uses powerful reasoning about rules and processes
 4. Technology that can be customized in its content and actions for individual organizations or people



Project Halo

- Vulcan R&D project to develop technology for the Digital Aristotle
 - Similar to an EC Integrated Project or a US DARPA project
 - Coordinate with the (few) other efforts in this area in the world
 - Find the best ideas/teams worldwide, and fund them to change the world
- Three major Project Halo thrusts
 - [Textbooks You Can Talk To](#) (Halobook)
 - New directions in knowledge representation (SILK)
 - Tractable nonmonotonic / higher-order rule systems
 - SME-based Knowledge Acquisition (AURA and SMW)
 - Address unscalable authorship costs for complex knowledge bases
 - Develop editorial processes and rules for knowledge authoring
 - Develop specific knowledge authoring technology for subject-matter experts



Project Halo and the Argument for Wikis

■ Three Project Halo challenges:

- Knowledge Formulation: Can you build the knowledge bases?
- Question Formulation: Can you query the knowledge bases?
- Question Answering and Explanation Generation: Can you get the answers?

■ Knowledge Formulation Challenge: create technology to build very large, computer-processable knowledge bases

- Millions of interlinked assertions, rules, and patterns, structured to support for question-answering algorithms, built and maintained in a cost-effective, reliable way
- Require solutions to AI issues that limited previous attempts at large-scale KBs
- Need cost-effective and scalable solutions

■ Wikis are one way to crowdsource logically simple information

- Highly reliable, Internet-scale, and incredibly cheap

■ Project Halo Goals for SMW

- Socially-driven low cost non-database knowledge acquisition for AI
- “Pay as you go” data integration via pages
- Emergent and evolving schemas



Vulcan's SMW support

■ 2007: A small dev team focused on improving SMW

- Bridge between an academic project and a Halo-viable piece of software
 - Markup tools that didn't require wikitext authoring
 - Coherence-promoting tools (auto-completion, gardening)
 - Visual query builder for ASK
 - Simple data structure browser (ontology browser)
- Provide support for Project Halo experiments
 - [Results were promising](#) for authorship; mixed for mapping
- Release as open source

■ 2011: One of the main players in a growing community

- Focus has been on requirements for the enterprise/commercial market
- Expanded forms, ontology management, Rules, Security, triplestores, notifications, visualizations, MS Office integration, gardening bots, external data APIs, core SMW work, scaling issues, etc.
- Active community forum (SMW Forum), testing program, documentation
- Six user group meetings to date
- Expanded SMW+ team: TeamMersion, AIFB @ KIT, Free University of Berlin, MES

■ SMW and its extensions are the software option for large-scale community-based knowledge acquisition



Developments since Spring 2010 SMWcon

■ SMW Community

- SMW+ downloads
 - ~2K/month SMW+ downloads from Sourceforge
 - Also at GitHub and at Google Code
 - Commercial support from Ontoprise
- SMW Forum traffic runs about 8K visitors/month, 5K uniques, busy discussion forum
- Vastly improved SEO (thanks to Wil Smith)
- Seattle-based SMW Forum mirror site with load balancing and caching

■ Last year's feedback-based developments in SMW+

- Semantic security: HaloACL backend
- Easier authoring:
 - Automatic semantic forms, better autocomplete, faceted browsing
 - Edit semantic data in the ontology browser
 - WYSIWYG editor based on CKedit
- External Data Support
 - Non-existent page handler for template-based access to database data
 - Data Integration via Ontostudio/Ontobroker
- Deployment framework , content bundles, refactoring, license revision

■ SMW has become a platform



2011 SMW+ Plans – Sustainment of SMW

■ Refactoring support for SMW-managed data

- Add/delete/modify multiple property values, instances, classes with a single action
- Consistently rename properties and categories throughout the wiki

■ Clean Up and Consolidate

- Multi-ontology management support
- Make Halo components extensible (esp. OntologyBrowser and SemanticToolbar) by defining explicit mechanisms to add functionality, implement sample extensions and document the API
- Integrated update mechanisms for scheduled tasks

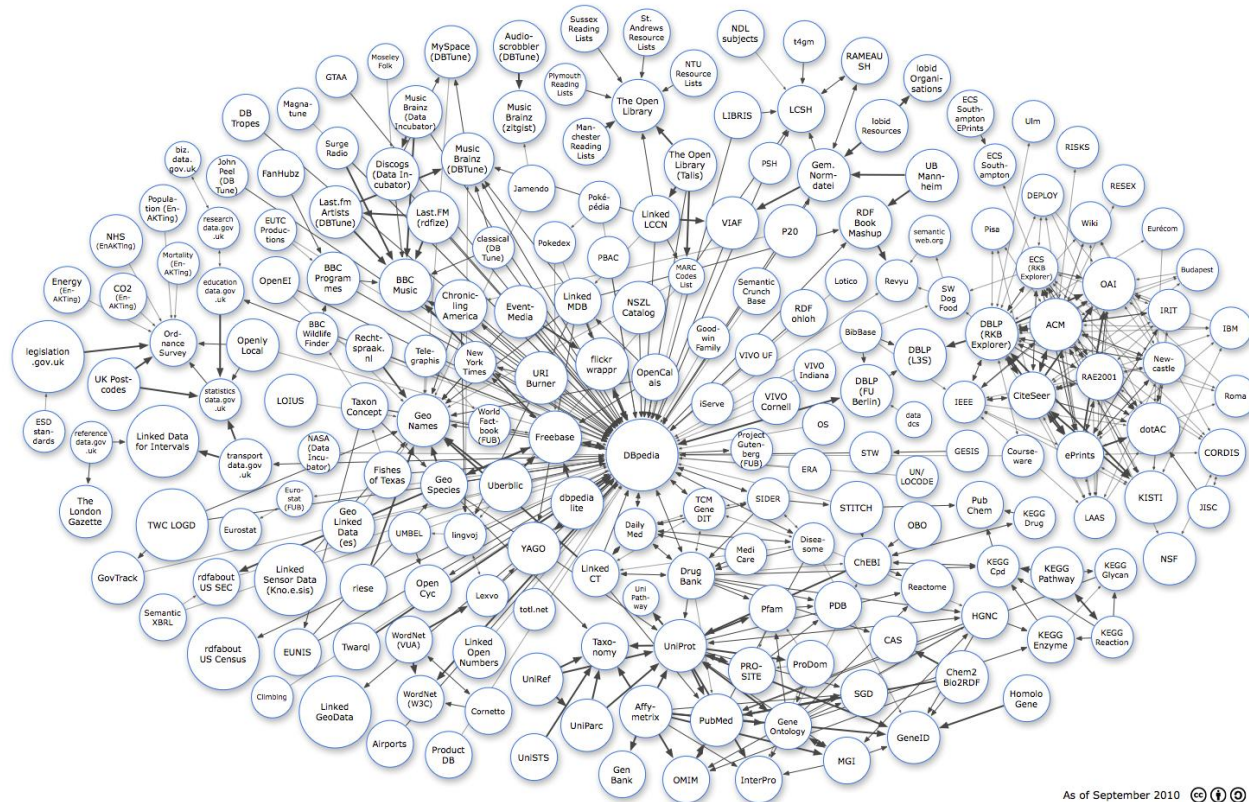
■ Improve Existing Features

- Improvements to query builder; add SPARQL support
- Improve Semantic Treeview usability
- Improve Deployment Framework
- Revise and consolidate some of the forms extensions

■ Community Support (docs, Forum, bugfixes, dissemination)



Major 2011 Thrust: SMW+ and Linked Data



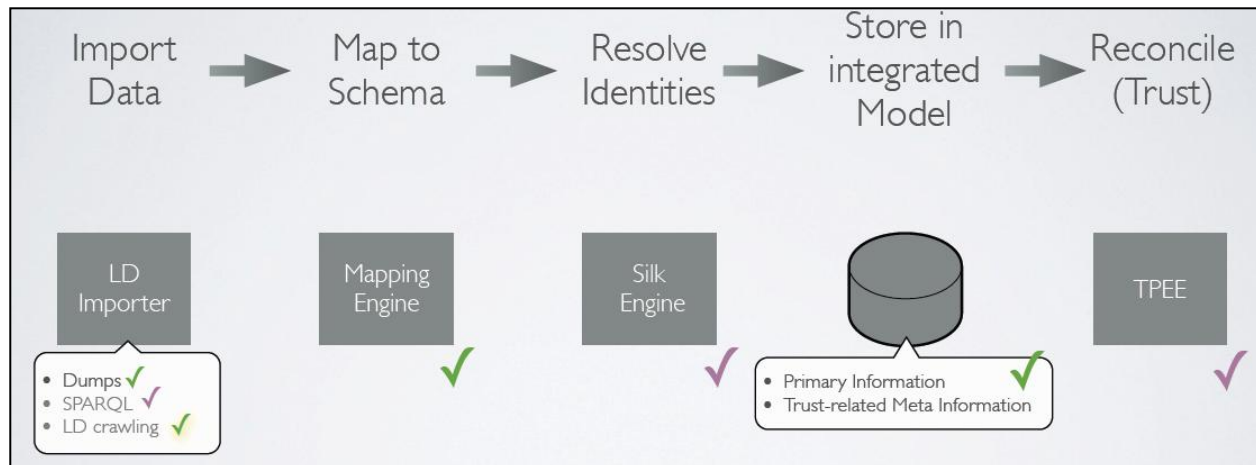
- SMW is positioned to be a browser platform, a curation platform, and a publication platform for Linked Data
- A commercially interesting role

Current Situation with SMW+ and Linked Data

■ ETL pipeline (SMW 1.5.2 + the Linked Data Extension)

- Load external data into a wiki-connected triplestore (Dumps or LDspider)
- Apply R2R mappings to map data source schema onto the wiki ontology
- Execute FU-Berlin SILK rules to apply identity resolution heuristics
- Store the resulting data locally
- Apply trust policies to determine orderings between data sources

■ Supports basic import/mapping of Linked Data



Linked Data Goals for SMW+ in 2011

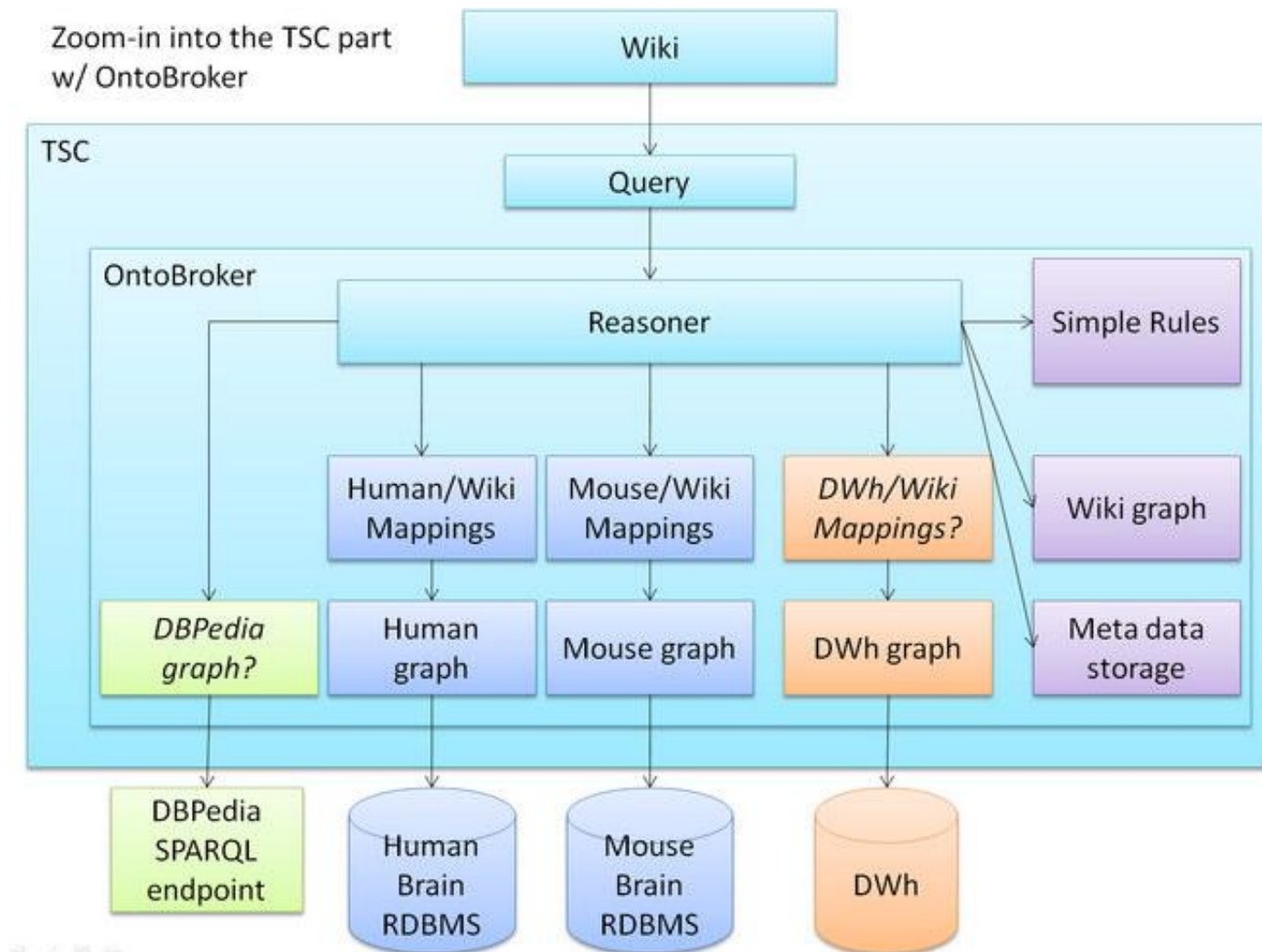
■ Improve current ETL-style pipeline

- Performance: Hadoop execution environment for SILK and R2R
 - A 100K biology set now takes ~19 hours on an m2.xlarge EC2 instance
- Transparency: Mappings, SILK rules, trust policies and data source definitions should be wiki elements
- Usability: UI for SILK and TPEE creation/edit in the wiki

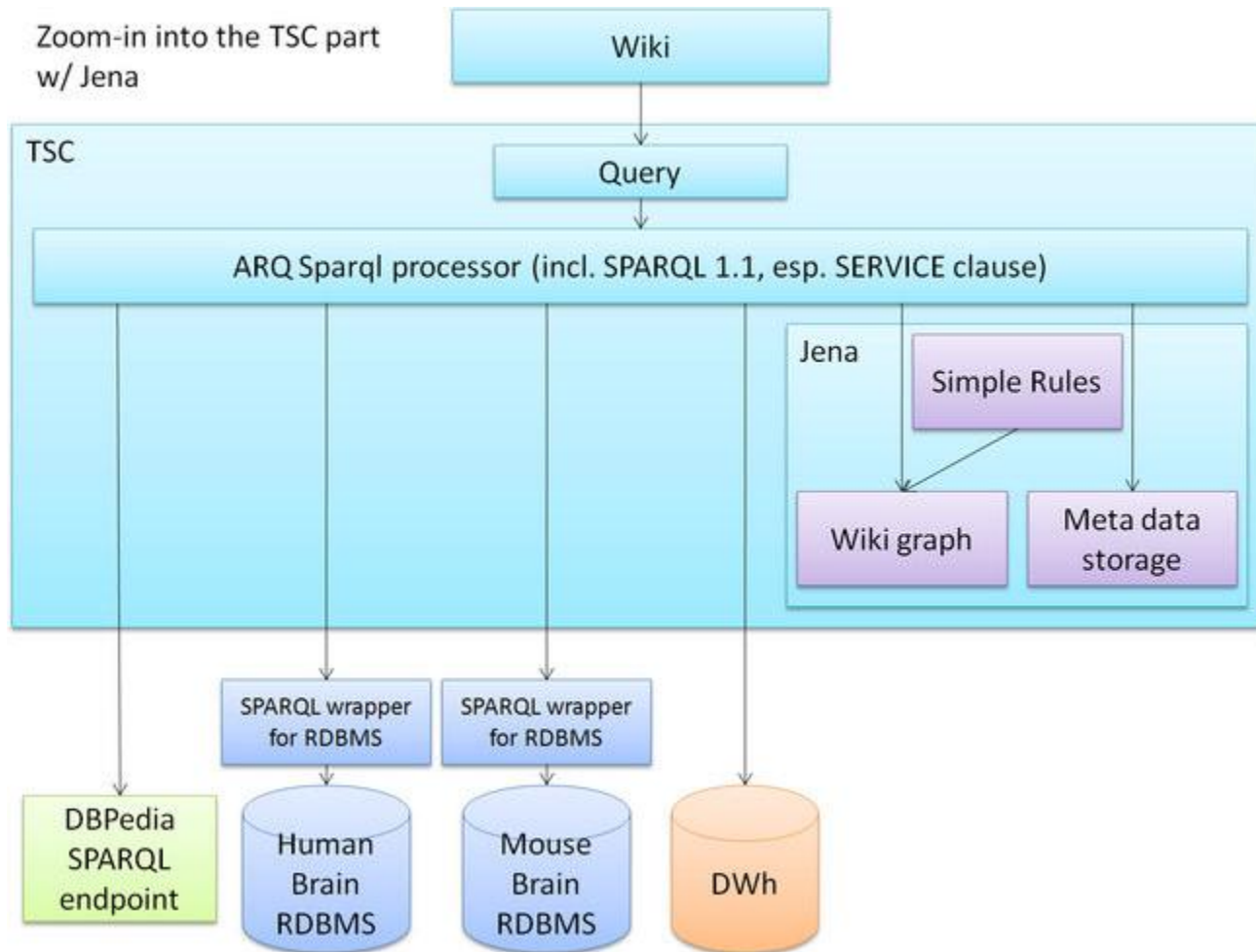
■ Access Linked Data without loading it

- Send SPARQL queries to any end point
- Perform federated queries over remote data sources
 - (Free) ARQ-based, using the SPARQL 1.1 SERVICE clause
 - (Licensed) Ontobroker based
- Query interface supports specifying remote SPARQL endpoints
- Aligns well with commercial requirements to play nicely with other data sources

Option 1: Ontobroker-based Federated Query



Option 2: ARQ-based Federated Query



Current SMW+ Programmatic – 4 Releases

The Usual Disclaimer: All Plans Are Subject to Change

■ Release 1: Mid-June 2011

- Multiple ontology support
- Improved Deployment Framework
- Definition of Linked Data artifacts in wiki markup
- SILK entity mapping editor

■ Release 2: Mid-September 2011

- SPARQL queries in Query Interface
- Data Access Projects (GUI and backend)
- GUI for modeling mappings between data sources/vocabularies
- Ontobroker Federated Queries: SMW+/TSC/OntoBroker executing queries according to a given "Data Access Project"



Current SMW+ Programmatics – 4 Releases

The Usual Disclaimer: All Plans Are Subject to Change

■ Release 3: Mid-December 2011

- Data Refactoring tools
- LOD artifacts available via enhanced retrieval, WYSIWYG, and/or Semantic Toolbar
- Remote SPARQL endpoints in the Query Interface
- Hadoop Execution Environment including workflow scheduler, Dump Loader, SPARQL Access and LDspider modules for Hadoop File System, R2R Mapping Engine (Hadoop Edition), Silk Engine (Hadoop Edition)

■ Release 4: Mid-February 2012

- API-based extension mechanism for basic SMW+ tools
- ARQ Federated Queries: Support for federated queries via #sparql-parser function and ARQ



■ Explore a wiki for all neuroscience data

- A collaborative framework for community-driven data integration and query
- Neuroscientists can contribute data sets and collaboratively edit metadata
 - Data mappings, entity resolution strategies, query ontologies/vocabularies, preference/trust orderings over data
 - Comment and rate all items
- Domain-appropriate query and visualizations

■ Status of initial prototype build

- SMW+ and Linked Data (BioRDF, Linked Open Drug Data, etc.)
 - Data Import, RDF Transformation, Ontology mapping, and Entity resolution
 - Basic datasource preference orderings
- 5 selected neuroscience data sources imported into SMW+
 - ABA, KEGG Gene, KEGG Pathway, PharmGKB, Uniprot
 - Triplestore built and queryable
- Basic wiki ontology created and mapped and being extensively revised
- Some specialty visualization components complete
- Sample Gene, disease, and pathway pages being created

■ We will evaluate the prototype in June and decide...

Sample Gene Data in SMW+

Set \$wgLogo to the URL path to your own logo image.

Change view | Markg | Log out

Search this wiki

Go

Search

My dashboardArticles and data

Main Page > Tacrine > Main Page > Query1 > Gene

CategoryDiscussion

RefreshHistoryEditMore

You have new messages (last change).

<http://myontology>

Queries for categories

Ask for all instances of "Gene" and for all instances of its subcategories

Schema information for category "Gene"

Properties

Property	Range/Type
AbaGeneld + ⓘ	String
Causes + ⓘ	Disease
EnsemblId + ⓘ	String
EntrezGeneld + ⓘ	Int
GeneSymbol + ⓘ	String
HGNClId + ⓘ	String
HPRDId + ⓘ	String
IMGTId + ⓘ	String
IsHomologTo + ⓘ	Gene
IsInvolvedIn + ⓘ	Pathway
IsTargetedBy + ⓘ	Drug
KeggGeneld + ⓘ	String
MgiMarkerAccessionId + ⓘ	String
NCBIGlId + ⓘ	String
NCBIGeneld + ⓘ	String
OMIMId + ⓘ	String
UniprotId + ⓘ	String

Properties whose range is "Gene"

Property	Range/Type
Involves + ⓘ	Gene
IsCausedBy + ⓘ	Gene
IsHomologTo + ⓘ	Gene

Categories: DiseaseOrGeneOrPathway | Top

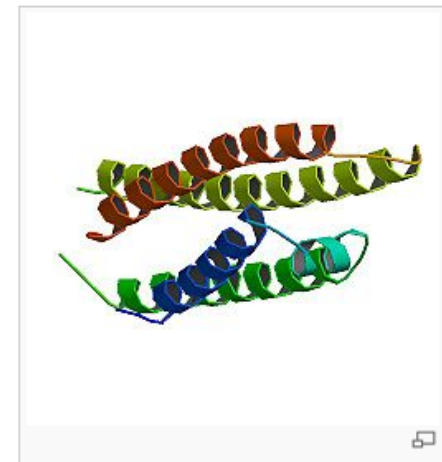
Mockup Gene Page (1)

Presenilin-1 is a **protein** that in humans is encoded by the *PSEN1* **gene**. Alzheimer's disease (AD) patients with an inherited form of the disease carry mutations in the presenilin proteins (PSEN1; PSEN2) or in the **amyloid precursor protein (APP)**. These disease-linked mutations result in increased production of the longer form of amyloid beta (main component of amyloid deposits found in AD brains). Presenilins are postulated to regulate APP processing through their effects on **gamma secretase**, an **enzyme** that cleaves APP. Also, it is thought that the presenilins are involved in the cleavage of the Notch receptor, such that they either directly regulate gamma secretase activity or themselves are protease enzymes. Multiple alternatively spliced **transcript** variants have been identified for this gene, the full-length natures of only some have been determined.[2]

Contents

[hide]

- [1 Function](#)
- [2 Gene](#)
- [3 Epistatic impact of APOE](#)
- [4 Alzheimer's Disease](#)
- [5 Interactive pathway map](#)
- [6 ***DATA***](#)
 - [6.1 List of Pathways \(and their genes\)](#)
 - [6.2 Related Drugs and Diseases](#)
 - [6.3 Expression Levels of this Gene in different Anatomic Structures](#)
 - [6.4 Publications about this Gene](#)



Mockup Gene Page (2)

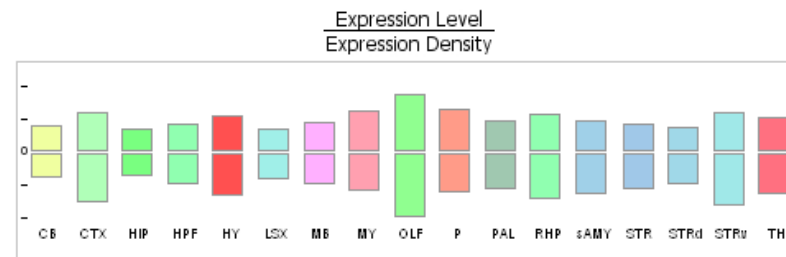
List of Pathways (and their genes)

	PSEN1	PSENEN	APH1A	APH2
Wnt signaling pathway	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Notch signaling pathway	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Neurotrophin signaling pathway	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Alzheimer's disease	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Related Drugs and Diseases

	Donepezil hydrochloride	Galantamine	Rivastigmine	Memantine
Alzheimer Disease	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Cardiomyopathy, Dilated	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Spondylothoracic Dysostosis	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Expression Levels of this Gene in different Anatomic Structures



from AIBS data of the mouse brain:

Publications about this Gene

1. Cummings JL, Frank JC, Cherry D, Kohatsu ND, Kemp B, Hewett L, Mittman B (2002). "Guidelines for managing Alzheimer's disease: Part I. Assessment". American Family Physician 65 (11): 2263–2272. PMID 12074525. <http://www.aafp.org/afp/20020601/2263.html>.
2. Cummings JL, Frank JC, Cherry D, Kohatsu ND, Kemp B, Hewett L, Mittman B (2002). "Guidelines for managing Alzheimer's disease: Part II.

Project Halo and the Digital Aristotle, Redux

- **Vulcan is committed to SMW's success**
 - Lowers the cost of knowledge authoring on the web
 - Brings an ever-increasing amount of the world's data online
 - Complements the curated, textbook-rooted knowledge in Halobook and other AI systems

- **“Requirements and Issues for SMW in Enterprise and Government” @ 3:15pm**
 - SMW in a multidatabase environment
 - Usability
 - User-level semantic authoring
 - Visualizations
 - Evolving ontologies
 - Deployability in the enterprise
 - Skins, samples, sandbox, web environment?
 - Is the current level of security sufficient?
 - Where is our competition (Sharepoint, Confluence) weakest?
 - Where is SMW weakest?



Requirements and Issues for SMW in Enterprise and Government

- SMW in a multi-datasource environment
 - Microsoft Office plugin experience?
- Usability
 - User-level authoring of ontology information – useful?
 - What is the next necessary visualization?
 - How do you manage ontology/data evolution
- Deployability in the enterprise
- Is the current level of security sufficient?
- Where is our competition (Sharepoint, Confluence) weakest?
- Where is SMW weakest?

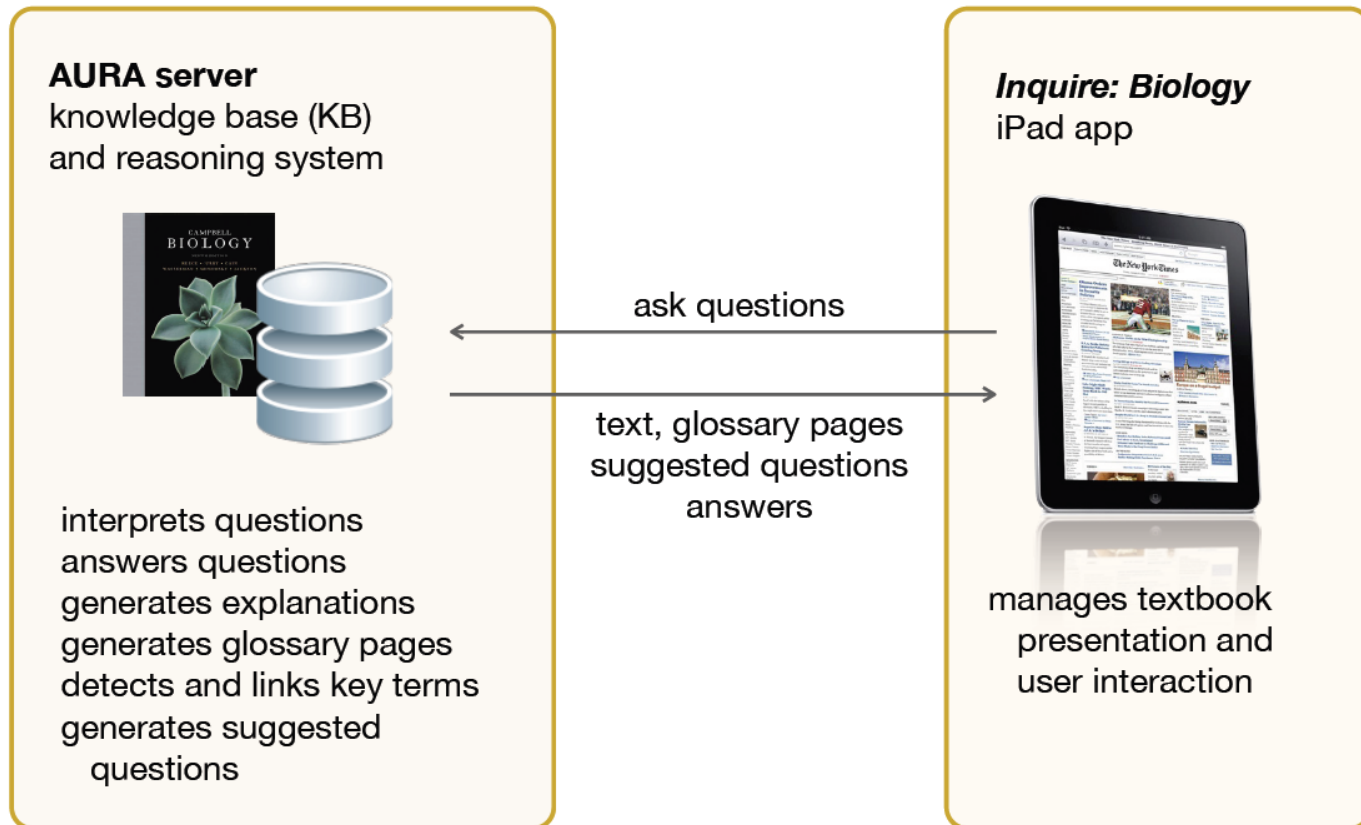


Thank You



The Halobook System

Two Main Components



Shown textbook:
Biology (9th Edition) by Neil A. Campbell and Jane B. Reece.
Copyright (c) 2011 by Pearson Education, Inc. Used by
permission of Pearson Education, Inc.

Inquire: Biology

6:26 PM 100%

7.4 Active transport uses energy to move solutes against their grad...

The Need for Energy in Active Transport

To pump a solute across a membrane against its gradient requires work; the cell must expend energy. Therefore, this type of membrane traffic is called **active transport**. The transport proteins that move solutes against their concentration gradients are all carrier proteins rather than channel proteins. This makes sense because when channel proteins are open, they merely allow solutes to diffuse down their concentration gradients rather than picking them up and transporting them against their gradients.

Active transport enables a cell to maintain internal concentrations of small solutes that differ from concentrations in its environment. For example, compared with its surroundings, an animal cell has a much higher concentration of potassium ions (K^+) and a much lower concentration of sodium ions (Na^+). The plasma membrane helps maintain these steep gradients by pumping Na^+ out of the cell and K^+ into the cell.

As in other types of cellular work, ATP supplies the energy for most active transport. One way ATP can power active transport is by transferring its terminal phosphate group directly to the transport

FIGURE 7.18 The sodium-potassium pump: a specific case of active transport.

1 Cytoplasmic Na^+ binds to the sodium-potassium pump. The affinity for Na^+ is high when the protein has this shape.

2 Na^+ binding stimulates phosphorylation by ATP.

3 Phosphorylation leads to a change in protein shape, reducing its affinity for Na^+ , which is released outside.

4 The new shape has a high affinity for K^+ , which binds on the extracellular side and triggers release of the phosphate group.

5 Loss of the phosphate group restores the pump's original shape, which has a lower affinity for K^+ .

6 K^+ is released, affinity for Na^+ is high again, and the cycle repeats.

FIGURE 7.19 Review: passive

Inquire BIOLOGY

The *Inquire: Biology* app contains the entire content of Campbell's *Biology*, a textbook used by many U.S. college undergrads and advanced high school students.

In portrait orientation, *Inquire* is optimized for reading, but also supports highlighting and note-taking.

Shown text and figures from:
Biology (9th Edition) by Neil A. Campbell and Jane B. Reece.
Copyright (c) 2011 by Pearson Education, Inc. Used by permission of Pearson Education, Inc.

A Glimpse Into the Knowledge Base

iPad 6:27 PM 100%

Protein

Definition Concept Map

Protein

A functional biological molecule consisting of one or more polypeptides and coiled into a specific three-dimensional structure.

Parts of Protein:

- Carbon Skeleton
- Hydrogen
- Carbon
- Functional Group
- Polypeptide

Kinds of Protein:

Actin, Allosteric Protein, Antibody, Antimicrobial Protein, Cadherin, Calmodulin, Chaperone protein, Chaperonin, *more...*

Tapping the "Concept Map" button takes the user to the corresponding concept map for protein.

Figure 5.15 An overview of protein functions.

Enzymatic proteins	Defensive proteins
Function: Catalyze biochemical reactions. Example: Digestive enzymes (amylase, lipase, and protease) break down food molecules.	Function: Protection against disease. Example: Antibodies (immunoglobulins) and white blood cells (leukocytes) fight off pathogens.
Storage proteins	Transport proteins
Function: Storage of amino acids. Example: Casein, the protein of milk, is the major source of amino acids for infants. Hemoglobin stores oxygen in red blood cells.	Function: Transport of substances. Example: Hemoglobin, the oxygen-carrying protein of red blood cells, transports oxygen from the lungs to other parts of the body.
Hormonal proteins	Receptor proteins
Function: Coordination of an organism's activities. Example: Insulin, a hormone secreted by the pancreas, causes other tissues to take up glucose, thus regulating blood sugar concentration.	Function: Response of cell to chemical stimuli. Example: Receptors bind with the molecules of a chemical signal, triggering molecular changes in other parts of the cell.
Contractile and motor proteins	Structural proteins
Function: Movement. Example: Myosin and actin are responsible for the contraction of skeletal muscles. Other contractile proteins are responsible for the movement of cilia and flagella.	Function: Support. Example: Collagen is the protein of hair, nails, tendons, and other skin components. It binds and holds cells and tissues together, providing structural support.

Showing text and figures from: Biology (9th Edition) by Neil A. Campbell and Jane B. Reece. Copyright (c) 2011 by Pearson Education, Inc. Used by permission.

JECT ALU

Using the Knowledge Base to Generate Concept Maps

The screenshot shows an iPad interface with a status bar at the top displaying 'iPad', signal strength, '6:27 PM', and '100%' battery. The app's title bar is 'Protein' with navigation icons on the left and 'QA', a settings gear, and a search icon on the right. Below the title bar, there are two tabs: 'Definition' and 'Concept Map', with the latter being selected. The main content area displays a concept map for 'Protein'. The map consists of yellow rectangular nodes. On the left, a vertical line labeled 'is a' connects the 'protein' node to a stack of four nodes: 'polymer', 'amphipathic molecule', and 'organic molecule'. To the right of the 'protein' node, a line labeled 'has parts' connects it to a stack of five nodes: 'carbon skeleton', 'hydrogen', 'carbon', 'functional group', and 'polypeptide'. A blue text box on the right side of the screen contains the following text:

The Concept Map for *Protein*, which provides an interactive graphical view of the partonomic and class information that is contained in the KB.

Each node is a link to its corresponding concept map.

14

JECT
ALU

Suggested Questions

6:27 PM

100%

7.4 Active transport uses energy to move solutes against their gradients

The Need for Energy in Active Transport

To pump a solute across a membrane against its gradient requires work; the cell must expend energy. Therefore, this type of membrane traffic is called **active transport**. The transport proteins that move solutes against their concentration gradients are called **active transport proteins**. This makes them different from **passive transport proteins**. They merely allow solutes to move down their concentration gradients rather than picking them up against their gradients.

Active transport enables a cell to maintain internal concentrations of small solutes that differ from concentrations in its environment. For example, compared with its surroundings, an animal cell has a much higher concentration of potassium ions (K^+) and a much lower concentration of sodium ions (Na^+). The plasma membrane helps maintain these steep gradients by pumping Na^+ out of the cell and K^+ into the cell.

Back in the textbook...

When the user makes a highlight—like the one above—*Inquire* presents a list of suggested questions based on the key concepts in the selection.

FIGURE 7.18 *The sodium-potassium pump: a specific case of active transport.*

Suggested Questions X

What is active transport?

What is the difference between active-transport and facilitated-diffusion?

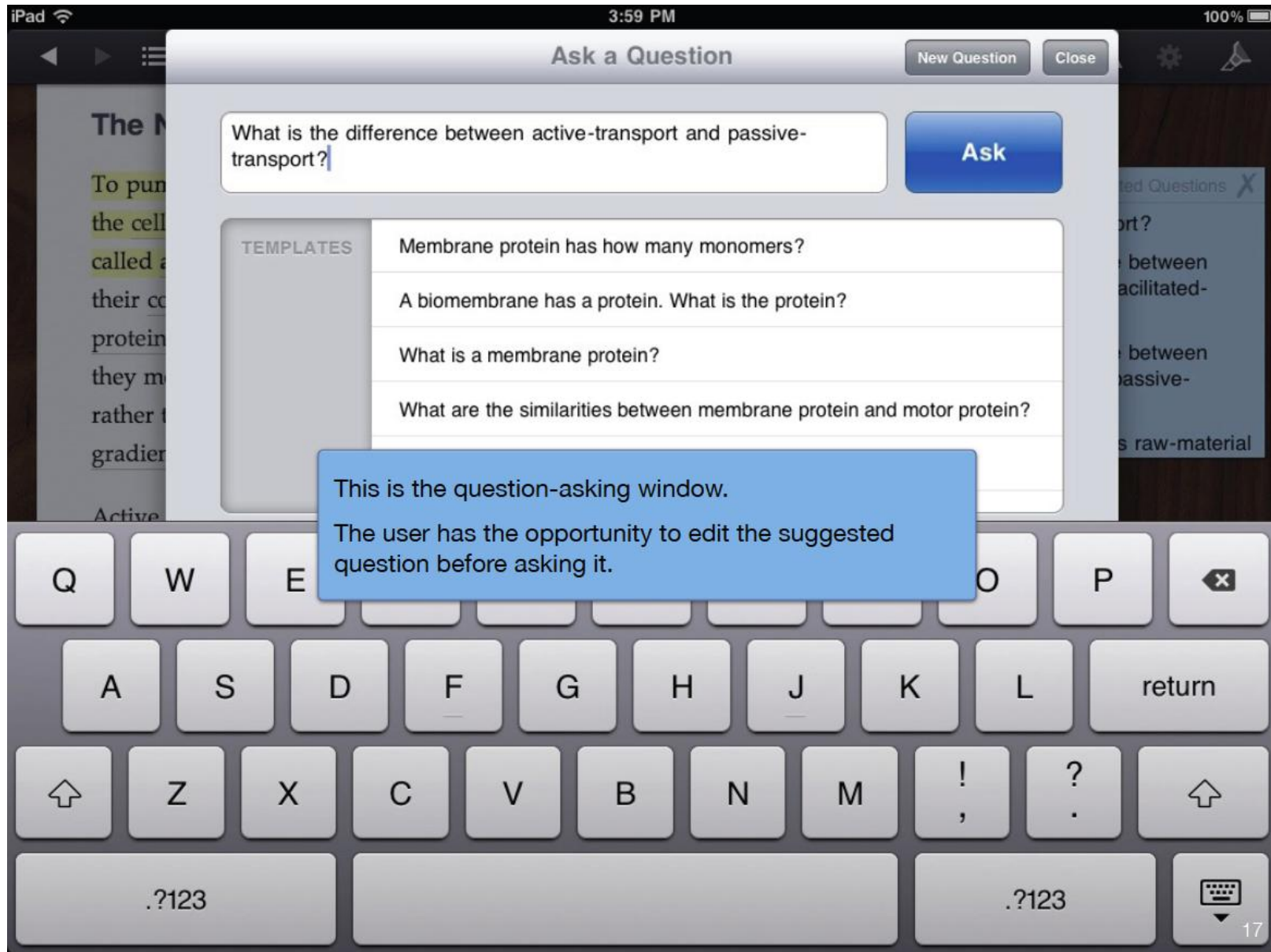
What is the difference between active-transport and passive-transport?

Is it true that energy is raw-material

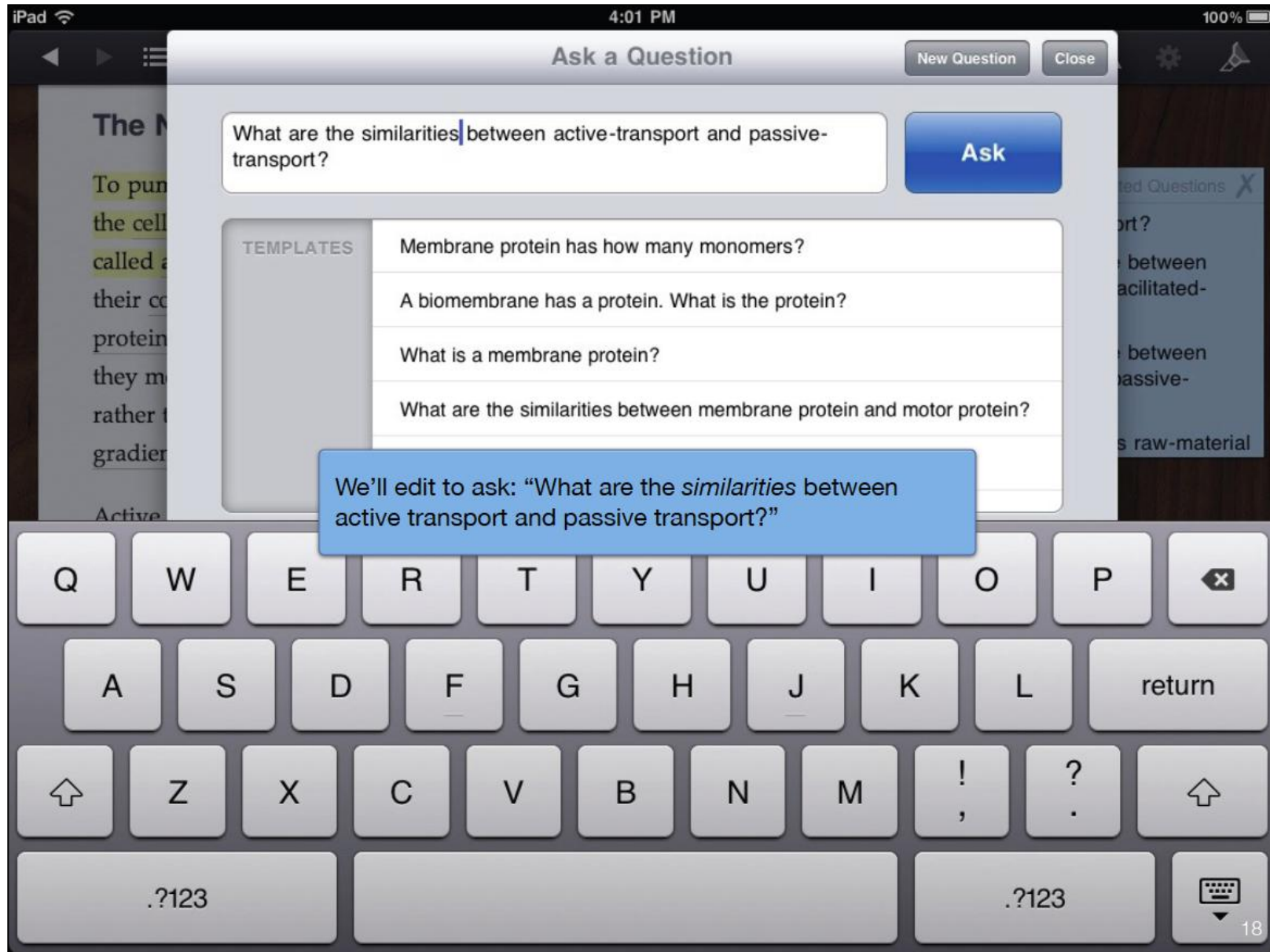
This is a note. It can be more than one line, and is entered using the iPad's virtual keyboard.

Showing text and figures from: Biology (9th Edition) by Neil A. Campbell and Jane B. Reece. Copyright (c) 2011 by Pearson Education, Inc. Used by permission of Pearson Education, Inc.

Asking Questions in Inquire



Editing Questions in Inquire



Answers from Inquire

iPad 4:02 PM 100%

Ask a Question

Answer

New Question

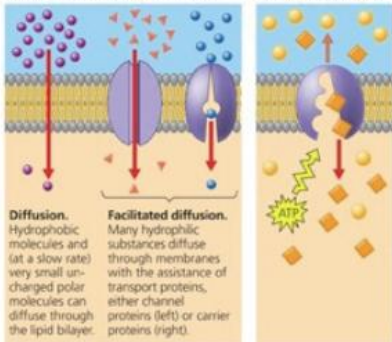
Close

What are the similarities between active-transport and passive-transport?

Active-Transport and Passive-Transport are similar in the following ways:

- They are both a kind of Event
- High Concentration Region is destination of Active Transport and Passive Transport
- Low Concentration Region is origin of Active Transport and Passive Transport

Passive transport. Substances diffuse spontaneously down their concentration gradients, crossing a membrane with no expenditure of energy by the cell. The rate of diffusion can be greatly increased by transport proteins in the membrane.



Diffusion. Hydrophobic molecules and (at a slow rate) very small uncharged polar molecules can diffuse through the lipid bilayer.

Facilitated diffusion. Many hydrophilic substances diffuse through membranes with the assistance of transport proteins, either channel proteins (left) or carrier proteins (right).

Active transport. Some transport proteins act as pumps, moving substances across a membrane against their concentration (or electrochemical) gradients. Energy for this work is usually supplied by ATP.

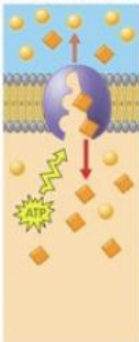


Figure 7.19 Review: passive and active transport.

AURA generates an answer, which includes a relevant image from the text. Links are automatically added for key terms.

Showing text and figures from:
Biology (9th Edition) by Neil A. Campbell and Jane B. Reece. Copyright (c) 2011 by Pearson Education, Inc. Used by permission of Pearson Education, Inc.

[BACK](#)

Cool Idea... But Does it Work?

■ User tests were performed in Chemistry

- 20 graduate students were each paid for 20 hours (over 1 month) to collaborate on semantic annotation for chemistry
- ~700 Wikipedia base articles
- US high-school AP exams were provided as content guidance

■ Initial Results (SMW+ 1.0)

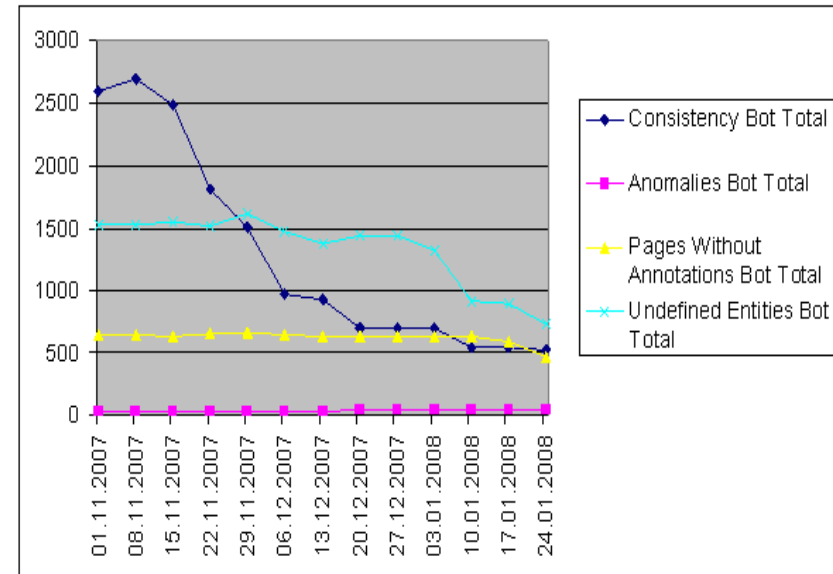
- Sparse: 1164 pages (entites), avg 5 assertions per entity
 - 226 Relations (1123 relation-statements) and 281 attributes (4721 attribute-statements)
- Many bizarre attributes and relations
- Very difficult to use with a reasoner

■ User testing and quality results for (SMW+ 1.1) extensions

- Initial SUS scoring (6 SMEs, AP science task) went from 43 to 61; final scores in the 70s
- 3 sessions using the Intrinsic Motivation Inventory (interest/value/usefulness); up 14%
- Aided by the consistency bot, users corrected 2072 errors (80% of those found) over 3 months

■ We have continued to build on this framework

Gardening Statistics for Test Wiki



Some Lessons Learned from SMW+ (and Freebase)

■ User Interface design matters

- This is core to MediaWiki's success
- Formal usability testing with SMEs matters a lot
- Zero-training matters a lot

■ Gardening matters

- Users need support for debugging
- Gardeners can do large scale ontology editing
- Supports “Schema Last” data engineering

■ User-created ontologies are not always well-designed

- Flatter than normal
- Cheaper than normal

[BACK](#)

■ Natural language is necessary to augment bare RDF(S) semantics

- Supplemental semantics can be usefully carried in natural language

